Molecular Dynamics forecasting of transmembrane Regions in GPRCs by Recurrent Neural Networks

Juan Manuel López-Correa
Computer Science Dept.
Univ. Politècnica de Catalunya
Barcelona, Spain
juan.manuel.lopez.correa@upc.edu

Caroline König
Computer Science Dept.
Univ. Politècnica de Catalunya
IDEAI-UPC - Research Center
Barcelona, Spain
ckonig@cs.upc.edu

Alfredo Vellido
Computer Science Dept.
Univ. Politècnica de Catalunya
IDEAI-UPC - Research Center
Barcelona, Spain
avellido@cs.upc.edu

***Abstract*—G protein-coupled receptors are a large super-family of cell membrane proteins that play an important physiological role as transmitters of extra-cellular signals. Signal transmission through the cell membrane depends on the conformational changes of the transmembrane region of the receptor and the investigation of the dynamics in these regions is therefore key. Molecular Dynamics (MD) simulations can provide information of the receptor conformational states at the atom level and machine learning (ML) methods can be useful for the analysis of these data. In this paper, Recurrent Neural Networks (RNNs) are used to evaluate whether the MD can be modeled focusing on the different regions of the receptor (intra-cellular, extra-cellular and each transmembrane regions (TM)). The best results, as measured by root-mean-square deviation (RMSD), are 0.1228 Å for TM4 of the 2rh1 (inactive state) and 0.1325 Å for TM4 of the 3p0g (active state), which are comparable to the state-of-the-art in non-dynamic 3-D predictions, showing the potential of the proposed approach.**

***Index Terms*—Deep Learning, Recurrent Neural Networks, LSTM, Molecular Dynamics, GPCRs**

I. Introduction

G protein-coupled receptors (GPCRs) are a large and diverse super-family of eukaryotic cell membrane proteins that play an important physiological role as transmitters of extra-cellular signal [1], making them relevant for pharmacology [2]. Around the 34 % of the drugs approved by the US Food and Drug Administration [1]. This has led, over the last decade, to active research in the field of proteomics. The functionality of proteins is determined by their 3-D structural configuration, but, to gain insights about the signal transmission mechanisms at the receptor, the conformational changes must be analyzed from a dynamical point of view. Computer-assisted MD simulations allow the study of the dynamic behavior of the receptors, particularly in the presence of drugs. Research on MD simulations has gathered pace in recent years, facilitated by the existence of MD repositories, such as the GPCRMD for MD simulations of GPCRs [1].

To understand the molecular basis of signal transmission, the transmembrane regions are the most important part of the receptor [3], as a conformational change is needed in these regions to transmit the signal trough the cell membrane. For this reason, the present study focuses on the analysis of the MD of the transmembrane regions of a GPCR receptor using ML approaches and, particularly, a variant of Recurrent Neural Networks (RNNs) [4], Long Short-Term Memory (LSTM) [5], which solves a limitations of the RNN architecture, namely the inability to learn information originated from far past in time. LSTMs are able to accumulate information for a long period of time by allowing the network to dynamically learn and to forget old aspects of information. Recent work has shown their potential to mimic trajectories produced by simulations [6], [7], achieving accurate predictions for short-term windows. To the best of our knowledge, there is no reported work on their use in the prediction of MD for the discriminating between the important regions of the of GPRCs, such as the extra-cellular, intra-cellular and transmembrane regions.

In previous research [8], we carried out a preliminary study of this kind using unidirectional and bidirectional LSTM. This study revealed the usefulness of LSTM to predict MD trajectories of a GPCR receptor as a whole. In the current study, the relative relevance of the different regions of the receptor is analyzed. A unidirectional LSTM is used to predict the MD of the different GPCR regions in two states of the beta-2 adrenergic receptor (β2AR). Simulations are analyzed for the 2rh1 structure (inactive state) and 3p0g structure (active state), both with full agonist (BI-167107).

II. Materials

A. GPCR MD simulations

The MD simulations used in this study were created in Google Exacycle cloud computing platform [9]. They comprise 10,000 trajectories of the β2AR-rh1 GPCR inactive (2rh1) and active (3p0g) receptor state with a full agonist. The receptor consists of 282 and 285 amino acids for inactive and active state respectively. Each trajectory describes the 3-D position of the receptor along 28 consecutive time-steps, which are hereon referred to as frames. The time elapsed between each frame is 500 pico-seconds.

B. Structural Sequence Domains

The GPCRs have three structural domains, namely a seven-helix transmembrane (TM) domain, an extra-cellular domain built by the N-terminus and three extra-cellular loops (EL) and the intra-cellular domain including the C-terminus and

© 2022 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes,creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. DOI: 10.1109/BHI56158.2022.9926945

three intra-cellular loops (IL) [10]. Table I provides a detailed description of each regions of the β2AR-GPCR receptor under study.

TABLE I: β2AR-GPCRs amino acid distribution by regions for inactive (2rh1) and active state (3p0g).

Region	State	amino acid id
N-terminus	3p0g	[∗ -23)
N-terminus	2rh1	[∗ -30)
TM 1	2rh1/3p0g	[30-60)
IL 1	2rh1/3p0g	[60-67)
TM 2	2rh1/3p0g	[67-96)
EL 1	2rh1/3p0g	[96-103)
TM 3	2rh1/3p0g	[103-136)
IL 2	2rh1/3p0g	[136-147)
TM 4	2rh1/3p0g	[147-171)
EL 2	2rh1/3p0g	[171-197)
TM 5	2rh1/3p0g	[197-229)
IL 3	2rh1/3p0g	[229-267)
TM 6	2rh1/3p0g	[267-298)
EL 3	2rh1/3p0g	[298-305)
TM 7	2rh1/3p0g	[305-328)
C-terminus	2rh1	[328- ∗)
C-terminus	3p0g	[342- ∗)

∗ *Unresolved loops by crystallography*

For 2rh1 structures entail residues 30-342, and for 3P0G residues 23-344. Both have gaps in the sequence, where the intra-cellular loop 3 (IL3) between TM2 and TM3 is replaced in 2rh1 and 3P0G with T4-lysozyme and a nanobody, respectively. These residues are 231-262 for 2rh1, and 228-264 for 3P0G. Since β2AR remains functional even in the absence these regions.

Figure 1 represents the common structure of a β2 adrenergic GPRC. The 7 TM, 3 IL and 3 EL regions are shown. In addition, ligand binding with the protein is displayed in an image inset.

III. Methods

A. Theoretical methodology

1) The Long Short-Term Memory model: LSTM [5] are neural networks of the RNN family that are designed for the analysis of temporal data. In short, LSTM has a input gate *(i)*, a forgetting gate *(f)*, one memory gate *(c)* and an output gate *(o)*. The input gate decides whether to let the incoming signal go through to the memory gate, or block it. The output gate could allow a new signal output or avoid it trough the memory gate. The forgetting cell is responsible to remember or to forget previous state of the memory gate. The update of memory gate states is carried out by feeding previous output gate to itself by recurrent connections of two consecutive time steps. The reading and writing memory cell is controlled by a group of sigmoid gate (*x*). At a given instance of time, the LSTM receives inputs from different sources: the current positions X_{xyz} as the input, the previous hidden state of all LSTM units (*h*) as well as the previous memory gate state $c_{(t-1)}$. Then, the output gate returns the probability of the next positions on the sequences (Px, Py, Pz). A schematic representation is shown in Figure 2.

In previous research [8], unidirectional LSTM (ULSTM) and bidirectional LSTM (BLSTM) were applied to predict the trajectories of the MD simulations of the receptor described in this study. ULSTM works by processing data in the forward direction, while BLSTM processes sequence data in both forward and backward directions with two separate hidden layers [11]. The best forecasting performances was obtained by ULST and, for this reason, they are the method of choice in the current study.

B. Experimental methodology

Data underwent linear max–min normalization [12] and was returned to the original range values in Angstrom (Å) units. The 3-D positions of amino acids were extracted for each frame. However, the original database included the positions of atoms instead of the position of amino acids. Therefore, the amino acids mass centers were calculated as the 3-D positions representing them. Two thousand trajectories per β2AR: 2rh1 (inactive state) and 3p0g (active state) both with full agonist (BI-167107) were used. We refer to trajectories as *nClones*. The LSTM training was carried out using the amino acid 3-D position (x, y, z) per frame. We refer to sequence lengths as *nSteps-in* and to length of the predicted sequences as *nSteps-out*, and the amino acid representative data point position as *center of the amino acids*. Values for these parameters were chosen according to the experiments reported in [8]. These were *nSteps-in* = 7, *nSteps-out* = 1 and *Center of the amino acids* = "center of the mass". The parameters of training configuration of the LSTM were: epochs=100, verbose=0, activation='relu', input-shape=(nSteps-in, length of amino acid chain). All other parameters of the Keras [13] framework were left by default. The data set was split in 5 folds with 400 *nClones* per fold. Four of them were used for training and cross-validation process, and the remaining fold was used for validation. Test predictions quality was assessed through *Mean Average Error* (MAE) [14].

In the experimental setup, three analyses were performed both for the *2rh1* and for the *3p0g* states. The first experiment evaluates the LSTM prediction error by seven TM, three EL and three IL regions. The second analysis focused specifically on the seven transmembrane sections, comparing between the *2rh1* and *3p0g* states. The same evaluation was carried out in the third analysis, but, in this case, focusing on the three intra-cellular and extra-cellular regions respectively.

IV. Results

The results are reported using two metrics with original range values in Angstroms (Å): MAE [14] for x, y, z coordinates and RMDS [15]. The standard deviation (std) [16] is also calculated for the RMSD metric. The prediction error values are shown for each region of the GPRC, such as TM, EL, IL. Table II presents the results for the *2rh1* state and Table III for the *3p0g* state. In both, the minimum error value is found in the 7TM region. These values were 0,1373 Å for the *2rh1* and 0,1521 Å for the *3p0g* state.

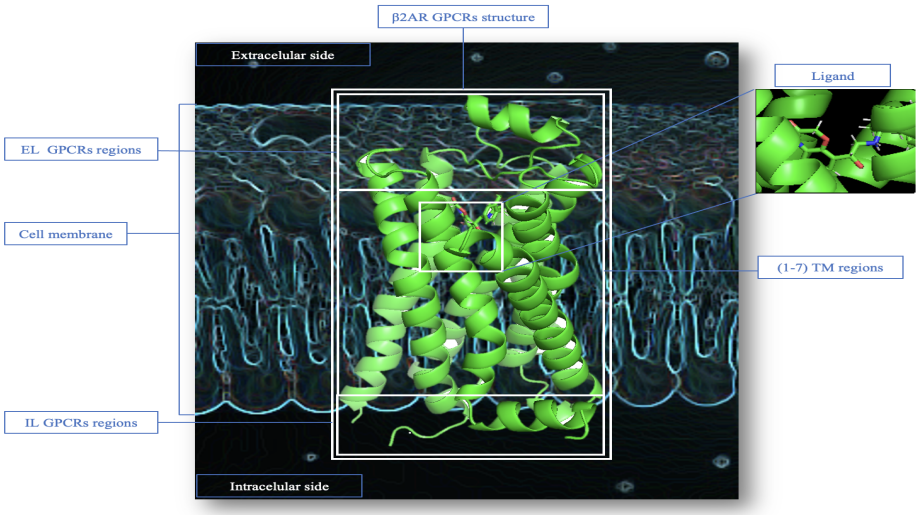

Fig. 1: Schematic representation of the β2 adrenergic GPRC. TM, EL and IL regions and Ligand binding with the receptor.

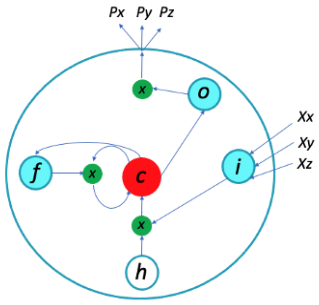

Fig. 2: Illustration of a LSTM unit.

TABLE II: Prediction error of the LSTM by MAE (*xyz*) and RMSD metrics. TM, EL, IL regions of the *2rh1* state.

Region	MAEx	MAEy	MAEz	RMSD
TM	0.0676	0.0731	0.0757	**0.1373** ± 0.022
EL	0.0589	0.0733	0.0808	0.1699 ± 0.022
IL	0.0802	0.0693	0.0816	0.2199 ± 0,037

The next experiment analyzes the prediction errors for each of the seven transmembrane regions. Table IV presents the prediction error per TM section (TM1-TM7) by RMSD value and its standard deviation. The best sequence predictions were found in TM4 section for both of the *2rh1* and the *3p0g* state with 0.1228 (Å) and 0.1325(Å) respectively, followed by TM3.

Focusing on the intra-cellular and extra-cellular regions

TABLE III: Prediction error of the LSTM by MAE (*xyz*) and RMSD metrics. TM, EL, IL regions of the *3p0g* state.

Region	MAEx	MAEy	MAEz	RMSD
TM	0,0759	0.0771	0.0784	**0.1521** ± 0.0359
EL	0.06932	0.0822	0.0877	0.1656 ± 0.0209
IL	0.0749	0.0768	0.0694	0.1711 ± 0,0256

Table V shows the same information for IL regions, while Table VI shows the RMSD and its standard deviation in *2rh1* and *3p0g* for EL regions. The minimum error was found for both states in IL1 section with 0,1719 (Å) for the *2rh1* and 0.1401 (Å) for the 3p0g state. In the case of the extra-cellular regions, EL1 had the minimum error with 0.1635 (Å) for the *2rh1* and 0.1361 (Å) for the *3p0g* state.

TABLE IV: Prediction error by RMSD metric by TM regions in *2rh1* and *3p0g* states.

TM section	2rh1	3p0g
TM1	0.1687 ± 0.020	0.1870 ± 0.053
TM2	0.1429 ± 0.022	0.1693 ± 0.065
TM3	**0.1234** ± 0.015	**0.1360** ± 0.027
TM4	**0.1228** ± 0.021	**0.1325** ± 0.025
TM5	0.1297 ± 0.014	0.1527 ± 0.032
TM6	0.1393 ± 0.032	0.1362 ± 0.013
TM7	0.1345 ± 0.029	0.1507 ± 0.034

TABLE V: Prediction error by RMSD metric by IL regions in *2rh1* and *3p0g* states.

IL section	2rh1	3p0g
IL1	**0,1719 ± 0.022**	**0.1401 ± 0.009**
IL2	0.1920 ± 0.035	0.1866 ± 0.038
IL3	0.2960 ± 0.055	0.1867 ± 0.028

TABLE VI: Prediction error by RMSD metric by EL regions in *2rh1* and *3p0g* states.

EL section	2rh1	3p0g
EL1	**0.1635 ± 0.019**	**0.1361 ± 0.008**
EL2	0.1678 ± 0.026	0.1754 ± 0.032
EL3	0.1758 ± 0.021	0.1851 ± 0.022

V. Discussion

The experimental results have shown that the LSTM performed best to predict the dynamics of the transmembrane regions followed by the extra-cellular regions. The intracellular regions yielded the highest prediction error. These results indicate that the regions dynamics of a GPCR are different, so that the ML model learns its prediction with different performance.

A more detailed analysis of the experimental results provided deeper insights about the MD of the specific receptor regions. Interestingly, for the transmembrane regions, TM3, TM4 showed the lowest error for both type of GPCRs states. This result indicates that the dynamics of those TM regions are the best predicted by LSTM; however, the std does not allow to conclude that there are significant differences between them. For the intra-cellular regions, the IL1 was identified as the region with the lowest prediction error in both simulations. Interestingly, the accurate prediction of the MD of IL1 region contrasts with the results for IL2 and IL3 with a significantly higher prediction error. In the case of the extra-cellular regions, for both simulations, the EL1 region was identified to have the lowest prediction error, while EL2 and EL3 regions yielded a significant higher prediction error. In addition, the experimental setup was carried out with ULSM, since the best forecasting performances was obtained in [8]. This may be given because the BLSTM trains in two direction, forward and backward over the sequence. However, the LSTMs only were evaluated for predict forward steps as ULSTM as do it in this work. LSTM also has drawbacks when it is used to process very long sequences. For this reason the next steps for the research line will provide more comparison studies/analyses with another Generative Models, such as, Transformers [17], Auto-encoder [18], etc.

VI. Conclusion

This study aimed to preliminary investigate the prediction ability of LSTM models of the different constituting regions of a GPCR receptor, in order to understand the problem complexity of MD modeling of GPCRs with ML models. It is important to discern the prediction capability of the ML model on the different regions, as the research on the signal passing mechanism can be focused specifically on a certain type of region. For example, the TM regions play an important role for the signal transmission because of their ability to change the shape of the transmembrane helices. It is known that an outer displacement of TM6 from the center of the helices and displacements of TM5 and TM7 are part of the activation mechanism of a receptor [3]. However, the research on the details of the mechanism of interaction between residues, which unchains the activation, is still ongoing. ML approaches, such as the LSTMs of this study, show promise for the in-detail analysis of the receptor dynamics and for the discovery of meaningful interactions between residues. The insights about the differences in the MD dynamics of GPCRs and the capability of the ML model to predict them should guide the design of future experiments to model specifically the MD of GPCRs with generative models, which have been successfully used for the modeling of protein MD [19].

Acknowledgment

This work is funded by Spanish PID2019-104551RB-I00 research project and by the PRE2020-092428 Ph.D. training program, through the Ministry Science and Innovation.

References

[1] Ismael Rodríguez-Espigares, Mariona Torrens-Fontanals, Johanna KS Tiemann, David Aranda-García, Juan Manuel Ramírez-Anguita, Tomasz Maciej Stepniewski, Nathalie Worp, Alejandro Varela-Rial, Adrián Morales-Pastor, Brian Medel-Lacruz, et al. Gpcrmd uncovers the dynamics of the 3d-gpcrome. *Nature Methods*, 17(8):777–787, 2020.

[2] Mathias Rask-Andersen, Markus Sällman Almén, and Helgi B Schiöth. Trends in the exploitation of novel drug targets. *Nature reviews Drug discovery*, 10(8):579–590, 2011.

[3] Naomi R Latorraca, AJ Venkatakrishnan, and Ron O Dror. Gpcr dynamics: structures in motion. *Chemical reviews*, 117(1):139–155, 2017.

[4] Zhe Wu, Anh Tran, David Rincon, and Panagiotis D Christofides. Machine-learning-based predictive control of nonlinear processes. part ii: Computational implementation. *AIChE Journal*, 65(11):e16734, 2019.

[5] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[6] Mohammad Javad Eslamibidgoli, Mehrdad Mokhtari, and Michael H Eikerling. Recurrent neural network-based model for accelerated trajectory analysis in aimd simulations. *arXiv preprint arXiv:1909.10124*, 2019.

[7] Sun-Ting Tsai, En-Jui Kuo, and Pratyush Tiwary. Learning molecular dynamics with simple language model built upon long short-term memory neural network. *Nature communications*, 11(1):1–11, 2020.

[8] López-Correa Juan Manuel, König Caroline, and Alfredo Vellido. Long Short-Term Memory to predict 3D Amino acids Positions in GPCR Molecular Dynamics, June 2022.

[9] Kai J Kohlhoff, Diwakar Shukla, Morgan Lawrenz, Gregory R Bowman, David E Konerding, Dan Belov, Russ B Altman, and Vijay S Pande. Cloud-based simulations on google exacycle reveal ligand modulation of gpcr activation pathways. *Nature chemistry*, 6(1):15–21, 2014.

[10] Caroline König, René Alquézar, Alfredo Vellido, and Jesús Giraldo. Systematic analysis of primary sequence domain segments for the discrimination between class c gpcr subtypes. *Interdisciplinary Sciences: Computational Life Sciences*, 10(1):43–52, 2018.

[11] Zhiyong Cui, Ruimin Ke, Ziyuan Pu, and Yinhai Wang. Deep bidirectional and unidirectional lstm recurrent neural network for network-wide traffic speed prediction. *arXiv preprint arXiv:1801.02143*, 2018.

[12] Ali Jahan and Kevin L Edwards. A state-of-the-art survey on the influence of normalization techniques in ranking: Improving the materials selection process in engineering design. *Materials & Design (1980-2015)*, 65:335–342, 2015.

[13] François Chollet et al. Keras https://keras. io. *Go to reference in article*, 2015.

[14] Tianfeng Chai and Roland R Draxler. Root mean square error (rmse) or mean absolute error (mae). *Geoscientific Model Development Discussions*, 7(1):1525–1534, 2014.

[15] Karen Sargsyan, Cédric Grauffel, and Carmay Lim. How molecular size impacts rmsd applications in molecular dynamics simulations. *Journal of chemical theory and computation*, 13(4):1518–1524, 2017.

[16] Dong Kyu Lee, Junyong In, and Sangseok Lee. Standard deviation and standard error of the mean. *Korean journal of anesthesiology*, 68(3):220, 2015.

[17] Francesco Giuliari, Irtiza Hasan, Marco Cristani, and Fabio Galasso. Transformer networks for trajectory forecasting. In *2020 25th international conference on pattern recognition (ICPR)*, pages 10335–10342. IEEE, 2021.

[18] Dor Bank, Noam Koenigstein, and Raja Giryes. Autoencoders. *arXiv preprint arXiv:2003.05991*, 2020.

[19] Matteo T Degiacomi. Coupling molecular dynamics and deep learning to mine protein conformational space. *Structure*, 27(6):1034–1040, 2019.